DOCUMENT RESUME

ED 091 426 TM 003 640

AUTHOR Ault, Leslie H.
TITLE Multiple-Choice Versus Created-Response Test Items.
PUB DATE [72]
NOTE 18p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Arithmetic; History; Intelligence Differences; Junior High School Students; *Multiple Choice Tests; *Response Mode; Science Tests; Sex Differences; Socioeconomic Status; *Test Construction; Test Reliability; Tests
IDENTIFIERS Free Response Test Items; *Test Items

ABSTRACT

The issue of multiple-choice (MC) vs. created-response (CR) test-item formats was reexamined at the eighth-grade level in three subject areas: general science, American history, and arithmetic. In each subject area, alternate forms with the same item-content but differing in which items were in which format were prepared from standardized tests. Between 269 and 289 students took each form. Measurement equivalence was substantiated by correlations corrected for attenuation between MC and CR items. Subgroups composed by sex, intelligence, and socioeconomic status (S.E.S.) showed no interactions with relative MC vs. CR discrimination, but one interaction was found with relative difficulty. In arithmetic, CR items were relatively more difficult (than MC items) for lower than for higher S.E.S. students. Comparison of overall item-test discrimination favored the CR items in arithmetic and history, but there was no difference in science. (Author)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRO
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRE
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

ED 091426

TM 003 640

MULTIPLE-CHOICE VERSUS CREATED-RESPONSE TEST ITEMS

ABSTRACT

The issue of multiple-choice (MC) vs. created-response (CR) test-item formats was reexamined at the eighth-grade level in three subject areas: General Science, American History, and Arithmetic. In each subject area, alternate froms with the same item-content but differing in which items were in which format were prepared from standardized tests. Between 269 and 289 students took each form. Measurement equivalence was substantiated by correlations corrected for attenuation between MC and CR items ranging from .90 to 1.04 (mean = .99). Subgroups composed by sex, intelligence, and scio-economic status (S.E.S.) showed no interactions with relative MC vs. CR discrimination, but one interaction was found with relative difficulty. In Arithmetic, CR items were relatively more difficult (than MC items) for lower than for higher S.E.S. students. Comparisons of overall item-test discrimination favored the CR items in Arithmetic and History, but there was no difference in Science.

MULTIPLE-CHOICE VERSUS CREATED-RESPONSE TEST ITEMS

LESLIE H. AULT[1]

Teachers College--Columbia University[2]

The widespread use of multiple-choice tests in America followed the success of the Army "Alpha" test during World War One. These "new-type" tests were attacked at the time and are still attacked now on grounds that they encourage superficial learning and dilute the educational process. Nevertheless, the acceptability of multiple-choice tests was established by numerous empirical studies investigating their psychometric properties during the 1920's and early 1930's. Ruch (1929) is a good source for descriptions of many of these early studies. A typical early study consisted of administering a set of items in open-ended or created-response format, and then on a later day administering the same items to the same examinees but in multiple-choice or true-false format, with the result that the "new-type" tests were found to have reliabilities about as high as the created-response test and to correlate highly with it. As Lindquist (1969) has pointed out, these technical justifications combined with mechanical scoring capability to establish the multiple-choice item as the dominant type, a development that ignored the probability that "every type of test exercise is superior to every other type for some specific purpose or purposes (p. 355)."

---

[1]The author is indebted to Dr. Elizabeth Hagen, under whose chairmanship the dissertation on which this article is based was developed.

[2]Now at Hostos Community College of the City University of New York.

Since the 1930's empirical studies of test-item types or formats have been relatively infrequent, and many pertinent studies had some other issue as their main purpose. A notable exception is a dissertation study by Cook (1955), who reported correlations corrected for attenuation of .95 to 1.00 between multiple-choice and open-ended versions of contemporary affairs items given to college freshmen. However, the results of some studies have been less clear-cut, including reports that American college students did relatively better on multiple-choice tests than did British students, who in turn did relatively better on essay tests (Vernon, 1962); of several low--as low as .22--correlations between arithmetic items from standard multiple-choice tests and open-ended counterparts given to fourth-graders (Williamson & Hopkins, 1967); and of higher reliability for an open-ended geometry test than for any of three multiple-choice versions (Owens, Hanna, and Coppedge, 1970). These reports provided indications that a further study might be worthwhile. In addition, a systemmatic study could employ methodological improvements (such factor analysis and one test to a subject) over the old studies. The present study was intended as a reexamination of the measurement properties of multiple-choice (MC) and created-response (CR) test-item formats.

The tests were at the junior-high level, where most of the items are suitable for translation into CR format and where there is a mix between straight factual items and one requiring more sophistication to answer. Tests at higher grade levels have many items unsuitable for translation into CR format, while tests at lower grade levels seem to have a preponderance of straight factual knowledge. Examples of items in each subject are given below.

Science

MC paired (Form R, #3): which of the following diseases is carried by mosquitos? / A Cancer / *B Malaria / C Heart disease / D Tuberculosis / E Pneumonia (p = .71, bis = .65)

CR paired (Form S, #5): What disease is commonly carried by mosquitos? (p = .71, bis = .65)

MC untranslated (Form R, #7; Form S, #7): Evaporation of water will take place fastest on a day which is / *A hot and dry. / B hot and moist. / C cold and dry. / D cold and moist. / E variable in moisture and temperature. (p = .75, .79; bis = .39, .48)

History

MC paired (Form W, #5): Patriots in the Revolutionary War received important financial and military aid from the / A Indians. / *B French. / C Loyalists. / D Russians. (p = .58, bis = .57)

CR paired (Form V, #5): From whom did the patriots in the Revolutionary War receive important financial and military aid? (p = .30, bis = .80)

MC untranslated (Form V, #1; Form W, #1): The development of communication was furthered by the inventions of all of the following men except / A Guglielmo Marconi / B Alexander G. Bell / C S.F.B. Morse / *D Elias Howe. (p = .36, .36; bis = .55, .46)

Arithmetic

MC paired (Form Y, #1): Jim cuts a 15.6-inch length of copper pipe into 6 equal lengths. How many inches long is each piece? / A .026 / B .26 / C 2 / *D 2.6 / E 15 (p = .74, bis = .53)

CR paired (Form X, #3): Jim cuts a 15.6-inch length of copper pipe into 6 equal lengths. How many inches long is each piece? (p = .49, bis = .74)

MC untranslated (Form X, #38; Form Y, #38): Which of the following products must be an odd number? / A 99,918 x 99,917 / B 99,918 x 99,921 / C 99,926 x 99,921 / D 99,926 x 99,926 / *E 99,929 x 99,933 (p = .46, .46; bis = .44, .55)

METHOD

Specially-made tests with MC and CR items were prepared in General Science, American History, and Arithmetic. The items were taken from the Educational Testing Service's Cooperative Tests, with many of the original MC items translated into CR format. In each subject area, two alternate forms were assembled with the same item-content but differing in which items were in which format. Thus each test form contained (a) some items in CR format appearing in MC format in the alternate form, (b) some items in MC format appearing in CR format in the alternate form, and (c) some items appearing in MC format in both forms. The items in the last group could not be translated into equivalent MC items, but were used as "anchor" items. On the presumption that CR items would take longer to answer than MC items, a few items (least desirable statistically) were dropped from the original test forms in order to maintain the same time limit for administrative purposes. Further details are given in Table 1.

TABLE 1

Item Categories, by Form

	Science		History		Arithmetic	
form:	R	S	V	W	X	Y
Item Categories						
(a) CR with MC pairs	16	16	18	18	16	16
(b) MC with CR pairs	16	16	18	18	16	16
(c) MC untranslated	18	18	24	24	13	13
Total items	50	50	60	60	45	45

The examinees were the entire eighth grade in a suburban New York school. The tests were administered on separate days for each subject area under the direction of the regular teachers. The tests were distributed with the alternate forms in alternating order during the normal class period with a 40-minute time limit. Most students took one test in each of three subject areas, but some took only two tests, some only one, and a few none, depending on their attendance pattern.

In addition, sex, age, intelligence, and socio-economic status (S.E.S.) were obtained for the students. Sex and age were supplied by the students on the cover of the test booklets. Intelligence test scores were obtained from the school records in the form of stanines on the Lorge-Thorndike, or from other test results in a few cases. S.E.S was based on father's occupation (with reference to father's education and mother's occupation and education where helpful) as supplied by the students on the test booklet and as listed in the school records. A three-level categorization was made using Blau and Duncan's (1967) table broken into thirds.

A summary of the numbers and characteristics of the samples by test form is shown in Table 2. Some of the differences were noticeably large, but none were statistically significant at the .01 level, permitting comparisons to be made across equivalent samples.

TABLE 2

Numbers and Characteristics of Examinees, by Form

form:	Science R	Science S	History V	History W	Arithmetic X	Arithmetic Y
Total Number of Examinees	289	274	284	276	276	269
Examinee Characteristics						
% Male	49.1	59.5	53.9	54.0	54.0	53.5
% Age 13	83.0	78.8	77.5	83.0	81.1	78.8
Intelligence: mean stanine	5.66	5.47	5.63	5.38	5.66	5.43
standard deviation	1.87	1.75	1.81	1.79	1.74	1.87
S.E.S.: 3. upper (%)	24.9	22.6	23.2	24.6	24.2	25.3
2. middle (%)	37.0	37.6	39.1	35.5	37.0	36.8
1. lower (%)	37.7	39.8	37.3	39.9	38.4	37.9

Note: Age, intelligence, or S.E.S. was not known for no more than two students per test form.

The students answered directly in the test booklets by circling the letter corresponding to their choice or by writing in a word, phrase, or number. The correct answers to the CR items were typically short and fairly concrete, making their scoring highly objective. The scoring was checked by comparison of codes given by two independent scorers for a sample of 20 tests for each form. After correction of a few scoring inconsistencies thus uncovered, the remaining "scoring error" on CR items amounted to 12 errors out of 2000 entries-tolerably low. The scoring error on the MC items was 5 errors out of 4200 entries, either transcription mistakes or hard-to-judge circles.

RESULTS

The analyses of the data were aimed at four main questions: (1) whether MC and CR items provide equivalent measurement, (2) how MC and CR items compare in item-test discrimination, and whether there are any differences among subgroups divided by sex, intelligence, and S.E.S. between MC and CR items in (3) difficulty and (4) item-test discrimination.

Measurement Equivalence of MC and CR Items

The simple and direct way to investigate measurement equivalence is to correlate scores on the MC and CR items. This was done within each test form, with the MC items divided into the "paired" and "untranslated" categories. As shown in Table 3, the six correlations between the CR and MC-paired subsets ranged between .66 and .80 raw, but between .90 and 1.04 (with a mean of .99) after correction for attenuation. In addition, more often than not the CR items correlated more highly with the MC-untranslated items than did the MC-paired items, thus providing no indication of differences between the formats. On the basis of the correlations, the MC and CR formats did provide equivalent measurement in this study.

The issue of measurement equivalence was also examined by factor analysis. For each test form, a principal components analysis with varimax rotation was performed on the matrix of tetrachoric correlations

TABLE 3

Item Subset Difficulty, Reliability, Intercorrelations, by Form

	Science		History		Arithmetic	
form:	R	S	V	W	X	Y
CR-paired:						
Mean difficulty	.39	.36	.26	.19	.39	.39
KR20 Reliability	.72	.71	.76	.69	.81	.78
MC-paired:						
Mean difficulty	.57	.60	.51	.50	.56	.51
KR20 Reliability	.73	.74	.66	.70	.74	.75
MC-untranslated:						
Mean difficulty	.53	.52	.39	.38	.46	.42
KR20 Reliability	.74	.77	.70	.70	.75	.80
Intercorrelations:						
CR-paired, MC-paired						
raw	.66	.69	.70	.71	.80	.80
corrected for attenuation	.90	.95	.98	1.02	1.02	1.04
CR-paired, MC untranslated						
raw	.72	.71	.64	.64	.74	.71
corrected for attenuation	.98	.92	.92	.89	.97	.95
MC-paired, MC-untranslated						
raw	.68	.70	.63	.62	.73	.73
corrected for attenuation	.92	.95	.88	.92	.90	.90

among items, in an attempt to identify possible format-related factors. For five of the six test forms, the second and third factors accounted for only 3-4% of the variance (the first factor is typically a strong factor associated with whatever the test is measuring) and showed no relationships with item format. On one of the History tests (Form W), the second factor accounted for 10.3-4% of the variance (the first factor is typically a strong factor associated with whatever the test is me measuring) and showed no relationships with item format. On one of the History tests (Form W), the second factor accounted for 10.3% of the variance and showed a marked relation with item format in both the unrotated and rotated structures, with the MC-paired items highly positive, the MC-untranslated items positive, and the CR items mostly negative in their loadings. The result on Form W is interesting but unconvincing as a valid format factor in view of the results on the other five test forms. There is reason to associate the factor with very difficult CR items, which occurred in greatest numbers on Form W and contributed most of the negative loadings.

Relative Discrimination of MC and CR Items

The relative discrimination of items in MC and CR formats was examined by comparing the item-test biserial and point-biserial correlations for each "item-pair" in its MC format and in its CR format. The summary of these comparisons for each subject area is shown in Table 4. In Science, there was very little difference

between the MC and CR formats in discrimination, using either the point-biserial or biserial correlations. In Arithmetic, on the other hand, both measures favored the CR format in discrimination. In History, the comparison using point-biserials showed no difference, but the use of the biserial correlation showed a substantial difference in favor of the CR format. This discrepancy resulted from the fact that many of the CR items in History proved to be very difficult; the point-biserial, unlike the biserial, is markedly affected by the proportion correct. These comparisons can also be judged roughly from the reliabilities shown in Table 3.

Subgroup Differences in MC vs. CR Difficulty and Discrimination

Despite overall uniformities, there is the possibility that different groups may perform differently as a function of item format. This was investigated for subgroups composed by sex, intelligence, and S.E.S. For this purpose, the students were divided into two roughly equal groups on intelligence by stanines 6 and above, and 5 and below. Item analyses were performed for each subgroup separately, and comparisons between them made using the difference in difficulty and in point-biserial correlation for the MC and CR formats of each item pair. There were no significant differences between subgroups in relative (MC vs. CR) discrimination, but there was one significant interaction in relative difficulty. On Arithmetic, the CR items were relatively more difficult (than the MC items) for lower S.E.S.

TABLE 4

MC vs. CR Item Discrimination Summary

(Entries are based on MC minus CR discrimination for each item-pair, using point-biserial and biserial item-test correlations)

	Number of items with difference of:					
	.12 or more	.0 to .11	-.0 to -.11	-.12 or more	mean difference	t
Science:						
point-biserial	6	11	10	5	.008	.39
biserial	7	8	8	9	-.017	.69
History:						
point-biserial	6	11	10	9	-.002	.09
biserial	4	6	5	21	-.096	3.86**
Arithmetic:						
point-biserial	3	9	12	8	-.043	2.11*
biserial	5	6	7	14	-.074	2.10*

* $P < .05$ ** $P < .01$

students than for higher S.E.S. students. A logical explanation is that this effect was related to the overall difference in discrimination in favor of Arithmetic CR items. However, S.E.S. correlated only in the .30's with the total score and also with intelligence, which in turn correlated .72 with the total score but showed a weaker and non-significant differential difficulty. Possible explanations are greater computational accuracy or greater tendency to check one's answer among higher S.E.S. children.

DISCUSSION

The present study supports the commonly-held notion that MC and CR items provide equivalent measurement. Where discrimination among examinees is the main purpose in testing, as where grades are to be assigned or for correlational studies, the evidence suggests that MC items can be used in place of CR items without disrupting what the test is supposed to measure. Such is not the case where an absolute rather than a relative standard is sought, as with a criterion test or where the concept of "process levels" is considered important.

The suggestion that CR items may provide better discrimination than their MC counterparts--at least in Arithmetic and American History--is important for measurement theory. The effect, of course, would be to improve test reliability by using CR items instead of MC items, which would be desirable if other things were equal. However,

there are also considerations of mass scoring and administrative time. Obviously scoring time becomes more important as a consideration and favors MC items as the numbers of examinees increase. Present mechanical scoring capabilities for certain types of CR answers, such as described by Lindquist (1969), are promising but unavailable for routine use. It would be useful for some future research in MC vs. CR comparisons to employ such machines and thus exert pressure for their continued development. Administrative time assumes importance in that CR items apparently require more time than do MC items. This extra time could also be used to add items to an MC test, thereby increasing reliability to perhaps the same level as provided by a CR test within the same testing time. In the present study, estimates indicate approximate equality in reliability for MC and CR items based on equal administrative time, but it is unknown whether the time-per-item could have been reduced somewhat without unduly affecting overall reliability. In the Owens, Coppedge, and Hanna (1970) study, administrative time was equal and the CR version was superior in reliability to any of the three MC versions. Further research in relative MC vs. CR discrimination should pay close attention to optimal administrative times, as well as examine the effects for other subject areas, age levels, testing settings, and types of tests.

REFERENCES

Blau, P. M. & Duncan, O. D., The American occupational structure. New York: John Wiley, 1967.

Cook, D. L., An investigation of three aspects of free-response and choice-type tests at the college level. Unpublished doctoral dissertation, Univ. of Iowa, 1955.

Lindquist, E. F., The impact of machines on educational measurement. In 68th N. S. S. E. yearbook, part II: educational evaluation, new roles, new means. Chicago: Univ. of Chicago Press, 1969. Pp. 351-359.

Owens, R. E., Hanna, G. S. & Coppedge, F. L., Comparison of multiple-choice tests using different distractor selection techniques. Journal of Educational Measurement, 1970, 7, 87-90.

Ruch, G. M., The objective or new-type examination. New York: Scott, Foresman, 1929.

Williamson, M. L. & Hopkins, K. D., The use of "none of these" vs. homogenous alternatives on multiple-choice tests: experimental reliability and validity comparisons. Journal of Educational Measurement, 1967, 53-58.